DOI:
10.1111/ejed.12273

Document Version
Peer reviewed version

The Valid Assessment of Learning – Who Should Be Responsible ?

Paul Black
School of Education, Community and Society
King's College London

Abstract

The main aim of this paper is to establish that to meet the requirements of validity of students reported learning achievements, schools must be given the main responsibility for the summative assessments of these achievements. It is argued that in order to meet this aim it is necessary that assessments by schools be developed to meet the requirements of validity. Therefore, the paper will first review the relationship between students' learning and the assessment practices of teachers and schools, arguing both that the terms formative and summative should be understood as the two ends of a spectrum of the functions of assessment, and that the different forms and uses of assessment practices within this spectrum may serve to enhance learning in various ways.

The paper will then describe ways in which assessments in schools have been developed to meet the requirements of validity. It will be argued that this synergy, between assessment practices, the aims of learning and the validity of assessment outcomes, can only be achieved by the sharing of responsibility, for assessing and reporting on these outcomes, between classroom teachers, their schools and state agencies. This argument will be considered briefly in the light of current policies and practices in a range of agencies.

1 INTRODUCTION

The **main aim** of this article is to argue that the need, to establish that teachers and their schools are preparing their students for life beyond their school-days, has to be met by requiring teachers themselves to both achieve this aim and to produce the evidence that they are doing so. This argument implies that other methods for meeting this aim, notably the use of tests designed and administered at state level to measure and thereby ensure the 'accountability' of schools are not as effective, and indeed that they both have damaging effects on the quality of students' learning and produce results that are invalid.

Many state policy makers believe that teachers cannot be given the responsibility for satisfying their demand for accountability, either because they do not have the skills to implement the required measurements, or because it is not possible to ensure that the results produced by teachers and schools are comparable and trustworthy across the state. There is evidence that this belief is partly justified, and if this is the case then the main aim of this article cannot be achieved. So a **secondary aim** of this article is to survey and describe the work that has been designed to overcome this weakness and establish in-school skills and procedures whereby the criteria for comparability between, and trust, in schools' own assessments can be secured.

In comparing the results of state-test accountability with this of school-based strategy, a key criterion will be whether the results on which a range of users, i.e. employers, or providers of post-school education, or parents and the students themselves, can be assured that the results of student's assessment can justify the inferences that they have make on the basis of these results, i.e that the results are **valid.** Thus, in the discussions below, the criterion of validity will be high-

lighted in comparisons between the outcomes of different was of producing end-of-school assessments.

In what follows, the aims of this article will be developed in three main sections. Section 2 will explore the secondary aim by surveying the ways in which various types of assessment may be integrated within pedagogy both to help students to tackle the tasks required by their school curriculum work and to develop their capacity and confidence in tackling more complex tasks in the future.

Section 3 will build on Section 2 to address the main aim by describing ways in which valid summative assessments by teachers, could be secured, with evidence from several national and state systems which illustrates the diversity of practices between those which obstruct the ideal which is our first aim and those which have worked to meet it.

Section 4 will conclude the argument by a brief consideration of the prospects of meeting the main aim on a wider scale by future work.

2 THE ROLES OF ASSESSEMENT IN PEDAGOGY

2.1 Models of pedagogy

Pedagogy is often used as an inclusive term to cover all aspects of teaching and learning with such adjectives as critical, conflict and liberatory, to highlight its various functions. Alexander (2008) emphasized the variety of the issues involved:

> *pedagogy is the act of teaching together with its attendant discourse of educational theories, values, evidence and justifications. It is what one needs to know, and the skills one needs to command, in order to make and justify the many different kinds of decision of which teaching is constituted. Curriculum is just one of its domains, albeit a central one.* (p. 47)

In its many definitions, such terms as teaching or instruction are used to specify one dimension of pedagogy (e.g. Hallam and Ireson, 1999). Different models may articulate the components differently, e.g. Bruner(1966) treats instruction as the over-arching concept, whilst he only mentions the term assessment in the context of research studies which monitor the outcomes of a particular curriculum. Alexander lists the core acts of teaching as task, activity, interaction and judgment (p.78), saying very little about assessment in the last of these.

2.2 The formative – summative spectrum

The meanings and the importance of assessment were discussed in detail in Black and Wiliam's 1998 article and in many other studies of its role and meaning. Diversity is inevitable because the term 'assessment' can apply to any production of evidence about the effects that an activity is producing: that evidence may then be used for a single purpose, or for several purposes. The evidence involved should be sought and collected in the light of the purpose, or purposes, for which it is to be used. In educational assessment, the purpose could be to check on students' understanding of a statement which a teacher has just made, because if it has been misunderstood such failing must be corrected before proceeding with work which is dependent on that understanding. In contrast with this short-term purpose, an informal test at the end of a few weeks of classroom work on a topic might serve the long-term purpose of giving the teacher, and the students, an overview of what has been achieved before work proceeds to a new topic.

However, such examples should be seen as two points in a spectrum of purposes: a piece of written homework in which each student had to draw upon the topics treated in last two lessons would be an intermediate point in that spectrum, whereas a formal summative test at the end of a semester or school term might be the terminal limit of that spectrum. In practice any assessment can serve both formative and summative purposes in a variety of combinations and can lead to a variety of types of feedback which function within the teacher's overall model of pedagogy.

Any such model may be construed, in the light of its effect in determining the teacher's activity, as encompassing three stages: it should be designed to achieve the overall aims of the teaching and to implement classroom activities which will serve to achieve these aims. There will be a world of difference between classroom work which aims to develop the ability to tackle complex tasks and work which is designed to secure rote learning. The classroom activities which will follow as a teacher works to implement these different aims will be radically different from one another.

To take the argument further, it is necessary to explore the variety of types of feedback which may be used when any teacher's plan is implemented in the classroom, and to link these to the ways in which each might serve the aims.

2.3 Assessment as an intrinsic part of classroom activity

The variety in types of assessment is related to the various timings involved. Immediate responses can happen in the classroom, when a teacher responds, to a question or a proposed answer from a student by an exchange involving only that student. Alternatively, a teacher can open up a whole class discussion calling for alternative answers or comments from other students before giving a more general response. In both, the choice and subsequent expression of a response is challenging because a decision about how best to help will have to be made very quickly, and in at least some cases will involve dealing with novel and unexpected suggestions.

Several factors will affect the range and level of students' responses. A teacher's plan for work on a topic should be designed to match the existing level of knowledge and understanding amongst the majority of the students. If the level assumed is too elementary, then the students' may need no more that a rapid revision of what is already understood; if it is too far ahead of the students' understanding, then no useful discussion may be possible. Of course, teachers should always know how far students have progressed in their work on a topic, but research studies on the topic of progression have shown that too little is known about the choice of optimum sequence in developing young learners' understanding of a topic.

Such research starts by mapping out the logical sequence of development of a set of concepts, relating each to the simpler concepts on which its understanding is built, and to the more advanced concepts which will need to draw upon that understanding. The second step is to express the proposed model in a set of questions, each matched to one of the component concepts in the model. This collection is then given to samples of students to check that the inter-dependencies, and different levels of difficulty which they imply, are confirmed by the students' different results (Black, Wilson et al, 2011). Such work has usually revealed many inconsistencies, so that the assumed model has to be revised, often over several cycles of modification and empirical verification. The outcome can be a map of the interlinking of many component concepts which can be used as a guide for teachers' planning. Such maps of progression have been produced by work in England (Johnson and Tymms, 2011), in Germany (Hadenfeldt et al. 2016) and in the USA (Morell et al. 2017). The results of these studies may not seem consistent with one another, partly because each is developed within the curriculum and the assessment system of its country or state, and partly because different studies have focussed on different curriculum components. However, they do show that a national or state curriculum can only provide a general, and partly hypothetical, basis on

which teachers might plan their progression sequence in promoting students' learning.

In implementing a plan for progression in their classrooms, adjusting their plans using formative assessment is often essential. Thus work on a topic might start by asking questions to encourage a discussion whereby the matching of the plan to students' progress in understanding can be checked. Two factors will affect the range and level of students' responses. A *first* factor is the type of demand of a question. For example, if a teacher were starting lessons in science about light, instead of asking the class for the laws of reflection and refraction, he or she might ask:

> *Which is the odd one out – piece of white paper, mirror, picture, television? Why?*

This is a very open question, framed with the intention to invite students to exchange a variety of ideas and experiences about light in order both to start them to think about it and to give the teacher a general overall view of their existing ideas and of the terminology which they naturally use. There is no 'right answer' to this question, but in arguing about alternative answers the students will reveal their existing ideas about the nature of light. A *second* factor is the time allowed for students to respond: if a thoughtful answer is expected, students may need time to think and to compose ways of expressing of expressing their thoughts. Such a question might help guide, and be followed up by, a more tightly focused question to lead into the development of the teaching plan.

The aim of such questions is to get pupils talking about the subject of the lesson. If such talking is to develop, the teacher has to encourage it: so to ignore a strange response, or to merely state that it is wrong, is not helpful – a far better response might be 'Why do you think that?', and then to accept any explanation and ask the class 'Does anyone else have a different idea?' The teacher's task here is a delicate one, for on the one hand students must be helped to understand the new ideas which the students find challenging whilst on the other hand the discussion must not be allowed to wander too far away from the main aim of the lesson i.e. the formative feed back must be guided by linking to the validity of the teacher's aims.

Teachers may not be able to anticipate some of the suggestions that students may propose, so any teacher's planned choice of response to students must be contingent. However, a response can be chosen to cut short a discussion or to open up a discussion, i.e. it can to give the 'right' answer or explore the students proposals by provoking further discussion, e.g. by asking other students to comment on the proposed answer or to suggest alternative answers. The latter choice may open up a dialogue with one or more students and may lead to several exchanges with or between the students. Examples of such exchanges are given in the book by Black et al. (2003) and in articles by Bell and Cowie (2001), Coffey et al. (2011) Chi (2009), Ruiz-Primo and Furtak (2007) and Harrison et al. (2017): this last source gives vignettes of classroom discussions in four European countries.

Such development of dialogue in which students are engaged ought to be seen as a fundamental contribution to their learning. This point is emphasized by Alexander:

> *Children, we now know, need to talk, and to experience a rich diet of spoken language, in order to think and to learn. Reading, writing and number may be acknowledged curriculum 'basics', but talk is arguably the true foundation of learning.* Alexander (2008, p.28)

The dialogue which occurs in classroom discussions is only one form of dialogic interaction between student and teacher. On a different time-scale, the feedback which a teacher provides on a student's written work can also be an opportunity for such interaction. Such feedback might be limited to a single response, or it might ask the student to re-write a part of the text, or to make specified additions to it. For this feedback mode, both teacher and student can have more time to

compose their contributions, and the interaction can be more directly aligned to the learning needs of the individual student. However, where feedback on written work also involves giving it an overall mark, such marking may inhibit learning. Feedback can have different types of effect on the development of students' views of themselves as learners: it may be resisted by students who have an overall aim of proving themselves and are looking to feedback to confirm a view of their intelligence or character, rather than as a means to develop their learning. Dweck (2000) argues that this choice is dependent on whether learners believe that their intelligence or their capacity to learn are fixed, so that the purpose of feedback is to confirm such beliefs, or that the feedback is a means to develop these features. As Dweck puts it:

> *There's another mindset in which these traits are not simply a hand you're dealt with and have to live with, always trying to convince yourself and others that you have a royal flush when you're secretly worried it's a pair of tens. In this mindset, the hand you're dealt with is just the starting point for development. This growth mindset is based on the belief that your basic qualities are things you can cultivate through your own efforts.* pp.6-7

In work where teachers tried to give only comments, including suggestions about how to improve, it was found that students worked more productively with these when no overall mark was given. In addition, teachers who kept records for each student of the comments they had made and of that student's responses thought that such records were a better guide for reporting on each student's progress than a set of marks (Butler,1988).

Work of this nature can develop learners' capacity both to engage in and learn from interactive dialogue and to reflect critically on the detailed outcomes of their own work and to take initiatives to improve it. However the thinking processes which underlie such self-regulated learning require subtle consideration (Black and Wiliam, 2009). Research into these processes has been reviewed by Greene and Azvedo (2007): they present some of their findings by using the following headings :

1 Identify task	*A Conditions (of learner and context)*
2 Planning a response	*B Operations to transform input and own data*
3 Enacting a strategy	*C Standards: criteria for self-appraisal*
4 Adapting: reviewing perhaps re-cycling	*D Evaluation*

The left-hand column sets out a sequence of steps which can lead to formulation of a response: the right-hand column proposes a meta-level of features which may affect a student's response. For example, a student, finding a question too obscure, yet being anxious to provide a response, will provide an answer to a related but easier question. In other cases, the student may express a valid answer in terms that might seem meaningless. Lighthall, (1988) gives the example of a student who, when asked to say what the term 'infinity' means, said "I think it's the back of a *Cream of Wheat* box". The teacher said "Don't be silly Billy"**.** In later discussion with a counsellor, Billy explained that he had noticed at breakfast that on a box of a cereal there was a picture of a man holding this same box, and this picture showed a man holding the same box, and so on. That student had suggested a thoughtful answer, but the teacher failed to consider that he might be making an honest attempt and did not ask him to explain it. In general, when a student makes what seems on the surface to be a poor, even unhelpful, response, a teacher must consider the possibility that the student may be trying to respond thoughtfully, or may have mis-understood the question.

The discussion above has focused on 'on-the-fly' feedback during classroom discussion. However, students' written work can also be an occasion for feedback, particularly if the teacher concentrates on giving formative comments on the text rather than on simply giving marks. Indeed, written work and its

assessment can be an occasion for dialogue, albeit in a different time-scale from that of classroom discussion: some teachers have also, in their written feedback, asked students to re-write some parts of their work to enhance its quality. Of course, this form of feedback is less demanding because the teacher has more time to select and express the optimum feedback. Of course, summative tests can be used in the same way as other writing tasks as occasions for feedback.

A further occasion for feedback arises in peer-group discussions in which students give feedback to one another, either as an episode which varies the pace and involvements of whole-class dialogue, or when pupils assess one another's written work and thereby develop their understanding of the criteria of quality by applying them to concrete examples (Black, 2017).

2.4 Feedback for developing the capability of students to learn

In classroom and related work, teachers have to weave together the variety of possible short and medium length activities, the needs and opportunities that may arise in those activities, and the use of various types of feedback, to form a tapestry which will secure the aims of their pedagogy, One long term aim which should be pursued in all of this work is to build the capability of every student to become an effective and independent learner. Perrenoud (1998), in his commentary on Black and Wiliam's 1998 article about assessment and feedback, emphasised the importance of this aim:

> *This [feedback] no longer seems to me, however, to be the central issue. It would seem more important to concentrate on the theoretical models of learning and its regulation and their implementation. These constitute the real systems of thought and action, in which feedback is only one element.* p. 86

Such 'systems of thought and action' have been one feature of the account in Section 2, both in the quote from Alexander emphasising the need for feedback to promote dialogue, and in the analysis of Green and Azvedo analysing the thought processes involved when one is composing a contribution to a discussion. Both of these illustrate the aim of enriching students' learning, but there are other ways to contribute to this aim.

Classroom dialogue involves interactions between students themselves as well as between individual students and their teacher. Indeed, in the examples of classroom dialogue referred to above, students are often arguing with one another. Teachers should encourage such argument, for the teacher is not the only learning resource in the classroom – students can also be resources for one another. However, whilst it is a common practice to ask students to discuss issues in small groups and to then exchange conclusions between groups, studies of such group work have shown that it is often ineffective.

Reports by Mercer et al. (2004) and Blatchford et al. (2006) have shown that it is both necessary and rewarding to train students to work effectively in groups. For example, in Mercer's work it was found that after this training such words as 'think', 'should' and 'because' occurred three times more frequently in the group discussion than they had previously, and that groups so trained gained higher scores in subsequent tests of the topics discussed than comparable groups who had not been trained.

Teachers have encouraged the use of students' group work both to encourage the dialogue that can arise from the immediate issues in classroom discussion and for the feedback that can develop from dialogue based on a group or on its members' written work (Black et al., 2003). Marking by students of one another's written work has been explored by some teachers. In one classroom, the teacher handed back to each student their written work after recording an assessment of it, but without any comments or marks written on it: the students then worked in small groups, reading one another's work and then discussing

the strengths and weaknesses of each piece, thereby helping each student to reflect on the differences in quality between their own work and that of others in the group. One teacher reflected on the value of such peer interaction as follows :

> *We regularly do peer marking—I find this very helpful indeed. A lot of misconceptions come to the fore and we then discuss these as we are going over the homework. I then go over the peer marking and talk to pupils individually as I go round the room.* Black et al., 2003, p. 50

Such work can contribute to students' development as learners in two ways. The first arises as students have to discuss the criteria by which one written piece might be judged in comparison with another: such discussion will help students to understand the aims of the work, and the criteria for quality, through consideration of specific examples. The second is that by engagement in such peer comparisons, students will develop their capacity to reflect on their own work and to realize their own strengths and weaknesses. The importance of this feature was expressed in Wood's 1998 study entitled 'How Children Think and Learn' :

> *Such encounters are the source of experiences which eventually create the 'inner dialogues' that form the process of mental self-regulation. Viewed in this way, learning is taking place on at least two levels: the child is learning about the task, developing 'local expertise'; and he is also learning how to structure his own learning and reasoning.* p.98

Peer assessment can be used in a variety of contexts. The written work considered above might have been a homework task. However, student answers to a test set at the end of the teaching of a topic can be assessed, by peer interaction, in the same way, as can more wide ranging summative assessment tasks. In such cases, both the formative and summative aspects of assessment are involved, each making its particular contribution to students' learning.

An example of the style of classroom work that is focused on giving students a leading role in developing their own learning is given in the recent publications describing the achievements of projects supported by the European Union to enhance the practices of 'inquiry learning', mainly in school science but also in mathematics and technology (Harrison, 2014). For science inquiry at secondary school level, a teacher might start a lesson by presenting a class with an everyday process or artefact and proposing a question. Examples might be to explore the question "How does spaghetti change when it's cooked?", or to "list the sources and effects of ultraviolet radiation". In a first stage, students are asked to discuss in small groups the different phenomena involved, using their prior experiences, and ways in which their observations and ideas might be explored or tested experimentally. The groups then exchange their ideas in a plenary discussion where each group presents its proposals and all can question both the value and practicability of them. In the light of that discussion, each group then revises its proposals, makes a plan to carry out a relevant experiment, and then works to select equipment and carry out their proposed experiment. Finally, at a plenary meeting groups report the results of their experiments and their reflections on how these support, or lead to revision of, their initial understanding.

Throughout this process, the teacher is serving as a guide. There is a delicate balance to be kept between intervening to encourage the development of the learning within the groups by questioning, perhaps correcting, some of their decisions or explanations, and interventions which will give advice in such explicit terms that it narrows the opportunities for students' learning (Harrison, 2014). The teacher's task is to steer rather than direct the students' inquiries. Accounts of several inquiries, including some actual recordings of classroom work, are available on project web-sites (SAILS 2016, ASSISTME, 2017). A more detailed account of the principles involved, the lessons learnt in, and the differences between, the work in the eight participant countries, is given in Dolin and Evans (2018). An account of how such principles can be embedded in primary school lessons is given in Fitzgerald and Gunstone (2012).

A common feature of the several links between forms of assessment and the development of students can be summarized by the following statement:

> *Talk vitally mediates the cognitive and cultural spaces between adult and child, among children themselves, between teacher and learner, between society and the individual, between what the child knows and understands and what he or she has yet to know and understand.*
>
> Alexander 2008, p.92

What emerges clearly is that assessment is an intrinsic a part of any pedagogy that aims to support the development of learning. Moreover, where small groups of students are involved in interactive dialogue with one another, they reveal and develop qualities that might not be evident in any other context. This issue was spelt out by Harlen :

> *Recognising that, in the company of other learners, students can exceed what they can understand and do alone, throws into doubt what is their 'true' level of performance. Is it the level of 'independent performance' or the level of 'assisted performance' in the social context? It has been argued that the level of performance when responding to assistance and the new tools provided by others gives a better assessment than administering tests of unassisted performance.*
>
> Harlen 2012, p.32

Thus, the aims of pedagogy should include activities which can empower learners' to make their own decisions in well-informed and thoughtful ways. One dimension of this aim is to prepare learners to meet real and complex tasks that society will increasingly encounter and which will cross the boundaries between different school subjects (Stanley et al.. 2009) .

3 CAN SUMMATIVE ASSESSMENTS ACHIEVE VALIDITY

3.1 Validity for what purpose ?

The discussion above has been about contexts in which teachers and schools are free to develop their students learning in the ways that they judge to be most effective for this purpose. However, more has to be said about the role of summative assessments where their results are used for decisions affecting the futures, outside the classroom, of those involved. One outstanding area of concern is the effects on teachers' work of the pressures of state tests which are used to judge their work. However, there are more issues involved than the problems specific to high-stakes testing. For example, in many of the years of schooling, teachers have responsibility for those regular, year-on-year, or more frequent, summative assessments on which decisions important for each student's future are taken - so summative assessments ought to be designed to inform such decisions.

The quality of assessment results must be discussed in terms of the two criteria – reliability and validity. One attraction of nationally set tests, with their structures and arrangements designed to limit the causes of variability in the marking of students' responses, is that they can achieve high reliability. However, this cannot be the main criterion – which is that for a specified test, those who will make decisions on the basis of that test's results need to know whether such decisions are well guided by those results, i.e. whether the result is a valid for their decisions . Crookes et al. (1996) analysed the determinants of validity in terms of a linked chain of the relevant components. Their account served to direct attention to such matters as the tasks used, their administration and scoring, the aggregation of scores, the evaluations and judgments which followed, and the actual impact of the results. In their model, reliability, understood

as the consistency of the outcomes that would be obtained from an assessment process if it were to be repeated, is seen as one of the links in the chain and not as its final or most important component.

The following authoritative definition of validity, was jointly formulated by three organization in the USA:

> *Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests . . . It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated.*
>
> American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 9

The chequered history which led to this 'trinitarian' definition, and the several problems which arise when it is applied in practice, are discussed in detail by Newton (2012). The validity issue has also been analysed in detail by Pellegrino et al. (2016), where the account of a school course in mathematics is illustrated by a critique of the validity of the sources of evidence used and by data on the reliability of reported results. For the immediate purposes of this article, the key issue is that different groups may use test results for different purposes so that these results will be interpreted and used in many different ways. If a particular group is interested, to mention two examples, in solving problems in mathematics, or problems in interpreting different types of historical documents, precision about the meanings of such terms as "problem solving" or "interpreting" will be essential. Any entity involved, e.g. "problem solving", is a "construct", a concept which which Cronbach (1971) defined as follows: "A construct is an intellectual device by means of which one construes events. It is a means of organizing experience into categories".

3.2 Validity of summative assessments in education

Discussion under this heading should aim to answer two questions, namely "Who are the users of the assessment results?" and "What are the inferences which they need to make on the basis of these results?" Each possible answer to the first of these questions has specific implications for the answer to the second, so the possible answers will be considered as linked pairs. In the following discussion, each answer will be considered in relation to two main contexts – that of in-school assessments and that of externally mandated high-stakes assessments.

The students themselves are an important group of users: in the short term they may be re-assured or warned by information which reviews their progress, in the long term their choice of options for the next stage of study – in their current school, or between higher or further education options, or for full-time employment – will be guided by their school's policy and by any external, summative assessments. Parents are a second group: they use both sets of results to inform the guidance they give to their children, and to inform or provoke questions which they wish to address to their children's teachers or to the school management.

A third group of users are the teachers themselves. Summative assessment results may provide essential information to teachers of each subject studied in their own classrooms, or to a group of teachers with similar classes, or to the head of year in a comprehensive review of each student's progress. Of importance here is the information that a teacher who will no longer teach a student group, e.g. at the end of a school year, will provide for any future teachers of the same group. Further and different types of user will be those in overall charge of the school, to keep them informed about the progress in each school year and also to guide appraisal of the quality of the work of individual teachers.

A fourth group are those outside the school. One set of users are those responsible for the education policy of a school or for all schools across a state, both to locate problem areas, to help inform the optimum use of resources, and perhaps to reward, or to review, the work of each school. A different set are those who will use assessment information in selecting students, either for higher or further education options, or for full-time employment.

There is a wide diversity of interests between these groups, and any scheme for summative assessment can only be a compromise across these interests. It is nevertheless clear that for some purposes, formal written examinations cannot suffice. This is obvious in the case of "performance" subjects. For example, for music, there could be a combination of observation of performance in playing an instrument together with a test on paper : it could hardly be acceptable if student could be said to be 'good' at music if that student was not able to play any instrument or to sing from a score. However, at the other extreme, competence in history would not be expected to include an ability to make history, i.e. to 'make things happen'.

Most school subjects lie between these two extremes. In English, for example, producing a coherent piece of prose which displayed either creative ability or ability to present a survey of a field of work, might be judged important, whilst to do either of these on a newly specified topic within the time limits and stress of a formal examination might fail to correspond to these construct definitions and thereby restrict the validity. More generally, for any method used for assessment purposes, any claims by its designers for its validity must be taken seriously and explained in any report of that assessment's results.

For the example mentioned above of the assessment of inquiry-oriented practical work in science, the problem in the system of high-stakes assessments in England is that the only evidence for students' ability to plan and carry out scientific investigations is that they can answer questions on formal written tests about such tasks, the only role of their teachers being to confirm that they have carried out a prescribed list of routine practical tasks. There is no research that supports a claim that this evidence is valid. Indeed, the two European Union projects mentioned above have shown that developing and assessing this ability requires sustained effort by teachers both to develop and guide classroom work on open-ended investigations and to master the skills required for making valid assessments of the outcomes. In addition, for students who have chosen, and been accepted for, university courses in the sciences, it has been reported that many turn out to be incapable, initially, of carrying out the laboratory work required. From their study of this issue, and in particular of the differences between the direct assessment of practical skills (DAPS) and the indirect assessment of them (IAPS) Abrahams et al. (2013) conclude that:

> *We believe, given the numbers of students involved and the potential higher costs of employing more DAPS, teachers should be directly involved in the direct assessment of practical work. We would recommend that a greater use of teachers should be made in the summative assessment of their students' practical work, accompanied by a robust moderation procedure.* p.247

However, any users of assessment results for science which included direct assessments by teachers would need assurance that the results of such assessments were trustworthy and that that they were comparable between schools.

3.3 Schools' own assessments – comparability and trust

The discussion below describes a small-scale project which illustrates the issues involved in helping teachers to make valid summative assessments of their students work. This was an exploratory project involving eighteen teachers, nine of English and nine of mathematics, from three schools. They worked

together in nine whole-day meetings, held once every five weeks over 30 months, with the project team. Full accounts of this project have been published elsewhere (Black, Harrison et al . 2010, 2011). The account given here is mainly based on transcripts of interviews with the teachers, and on their written reflections on the work. A project with some similar aims has been described by Shavelson et al. (2008), but their aim was to build assessment items into an established curriculum in order to enhance teachers work, particularly in enhancing declarative and procedural skills in their classrooms.The three teachers who attempted to do this found this very challenging. They could not achieve the expected changes in their classroom teaching without sustained training which the project had not planned to provide.

This project's plan was to proceed in three stages. In the **first** stage the intention was for teachers to survey the assessment tools and practices they were using for their own summative assessments. Then in a **second** stage their understanding of validity, and of reliability, would be challenged in the light of an audit of the quality of their assessments, leading thereby to development of a shared formulation of their criteria of quality. This would lead to a **third** stage in which the teachers would develop and share methods for improving the quality of their summative assessments. This step would provide evidence of how understanding of validity was a guide to assessment practices, through the exemplification of the concept in classroom activities, and through a process of inter-school moderation.

The finding in the first stage were, rather like those of the Shavelson et al. project, discouraging . There were variations in practice, both between and within schools in each subject, and many teachers met a school requirement for end-of-year assessments by using sets of questions taken from previous national tests, or from other sources available on various web-pages, without any debates within each school to explore the qualities of these tests or their relevance to the aims of their teaching. The schools own summative tests were strongly influenced by the external tests, and because it was easy to copy items from these tests the teachers had not developed their own skills at composing items to reflect and reinforce their aims.

The team therefore cut short the first stage and proceeded immediately to the second by asking the teachers to discuss, in inter-school groups for each subject, their answers to the question "what does it mean to be good at your subject ?" When they had reached some agreement about this issue, they were then asked to consider whether or not their summative assessment provided valid evidence about whether, and to what extent, their students were 'good at' their subject . This produced very positive engagement. As one teacher of English reported at the end of the project :

> *The project made me think more critically about what exactly I was assessing. The first question I remember being asked ('what does it mean to be good at English?') gave me a different perspective on assessment. I find myself continually returning to this question.*
>
> Black et al., 2010 p.222

As teachers reported their findings, the research team challenged their understanding of validity in order to help them to formulate their own criteria of quality and to re-design their in-school assessments accordingly. All three schools worked on the assessments of students in year 8 (ages 12–13), both to limit the demands of the work and to carry out trials in a school year for which there were no externally imposed state assessments.

In then proceeding to the third stage, the King's team proposed that the schools' basic assessment evidence should be a portfolio of each student's work, on the basis of which the teachers could develop a process of inter-school moderation. The teachers had to design the contents of the portfolios, with appropriate achievement criteria, to work out how best to aim for these criteria, and to agree upon the procedures for assessing them. Some tasks were invented by the teachers themselves, others were

suggested by the team – drawing on published sources. All proposals were refined by the teachers through an iterative process, for the limitations of some only became evident when they were tried in practice. In this process, teachers were guided by the discussions which had helped them develop their own understanding of validity. For some proposals, students' work showed that the task was beyond almost all of them, whilst some other tasks were found so easy that they failed to challenge pupils or to discriminate between their different needs. Tasks had to be selected or adjusted to lie within this range; in their reflections at the end of the project, the teachers stated that this process of task adjustment was of particular value, both in developing useful material and in helping them to become better judges of task quality.

For teachers of English their main source of evidence was a portfolio containing samples of work that demonstrated each student's capability at the end of each unit, typically every half term. In this practice they drew upon well-founded models of course-work assessment (Smith 1978). The portfolio tasks were to cover work from the three strands, Writing, Reading, and Speaking and Listening, with three assessments for each strand. At least two of these would overlap in the aspects that they assessed, and one would be completed in controlled conditions. However, these teachers agreed that they could not, because of the project's limited time, explore some issues, notably the assessment of speaking and listening, and the matching of tasks to the interests and abilities of different students.

The mathematics teachers found it more difficult to design and implement portfolio tasks as part of their normal classroom work. Use of normal (i.e. pre-existing) classroom work turned out to be unsatisfactory because all the students produced much the same work in class. New types of task were required , and whilst this was welcomed, building these into regular teaching was very demanding. They also found the adoption of a holistic approach in their assessment difficult and unfamiliar (Brown, 1992; Wiliam, 1998).

Because the portfolios of tasks included many which were carried as a part of normal classroom work, and not constrained by the rules of formal tests, it was necessary also to discuss the ways in which suggested tasks were implemented. One teacher described this problem :

> *. . . I remember [colleague] and I doing the same task and obviously introducing it very differently and getting very different results. So I think having some agreed starting point is essential.* Mathematics teacher 2011 p.456

Other issues were: that students might not understand what to do when the task was of a type that they had not previously experienced: that teachers had to decide the extent to which students were to work together: and that teachers had to resolve tensions between giving students advice to help the improve their work whilst also achieving fairness in assessing the final product.

One approach was for teachers to give students formative guidance at all stages of their task work, thereby integrating such work into the normal sequence of teaching, whilst for the assessed work each student would be required, either to complete their own report on that task, or to tackle a similar, but novel, task individually but with whatever resources they chose to use. Poehner and Lantolf (2005) argued that students' attainments on a task after, rather than before, help has been given to improve their work produces a more valid assessment, but the project described here was not totally committed to this approach.

The overall outcome of this work was that teachers assessed each of their students on the basis of a portfolio, a collection of pieces of that student's work which would include both marks from formal tests and assessment of a variety of types of open-ended work, chosen so that between them the tasks met the requirement that the subject's assessment provided valid guidance to users of its results. To assist the

comparisons between the portfolio assessments of different schools, there had to be some comparability between the tasks results included in the portfolios. The teachers agreed beforehand that in each portfolio, half of the tasks would be the same across the three schools, and the other half could differ according to each teacher's choice.

It was then essential to set up ways to ensure that results from the different schools were both comparable in standards and trustworthy i.e. to reveal any inconsistencies between the teachers in their interpretations of the assessment criteria, and to ensure that assessments were free from personal bias or from uneven levels of help given to students. Such checking was essential, partly because it was found that student's reports on open-ended tasks were more difficult to assess than test papers. For this purpose the project arranged formal moderation meetings. Each school had to submit, before such a meeting, three samples of their assessed portfolios, one each from the top third, the middle third, and the bottom third of their results. These samples were circulated, with no indication on them of the teacher's assessment, between all three schools, and all would make and record privately their own assessments of them.

After this 'blind marking' process, the teachers would come to a moderation meeting, at which the independent assessments of the circulated samples by each of the group were tabled. A debate would then ensue to resolve any inconsistencies between the different assessments of the same samples. These meetings exposed some surprising levels of disagreement: in some cases there had to be exploration of differences in expected standards for different types of work, in other cases it emerged that a teacher had allowed for potentially valid features which could not be evident in the work itself. The fears often expressed in the literature – that plagiarism by some students might give them an unfair advantage – did not arise : the teachers were confident that they could detect such effects in their own students' work. Possible effects of personal bias were offset by the blind marking procedure and by the need for each teacher to defend any aberrant judgments in these meetings.

Initially it was feared that the time and effort involved in preparing and conducting moderation meetings would be seen by teachers as an unacceptable addition to their existing work-load. However, the following quotations from the writing of two of the teachers show that these fears were unfounded:

> *. . . that the moderation and standardisation process was incredibly valuable in ensuring rigour, consistency and confidence with our approach to assessment; that teachers in school were highly motivated by being involved in the process that would impact on the achievement of students in their classes.* English teacher, Black et al. 2011 p.459

> *And we've had moderation meetings, we were together with the other schools, teachers in other schools looked at how rigorous our assessment would be and they criticized what, you know, our marking criteria are. And we changed it, which has all been very positive.*
> Mathematics teacher, Black et al. 2011 p.459

The overall judgment of the teachers was that the work was both feasible and rewarding. They felt also that extensive professional training would be essential if all teachers were to understand the guiding principles and develop in their own work the practices required. The summing up on this aspect was expressed as follows by two of the group:

> *I think the department will need to go through the sort of thing that we've gone through, but obviously a little bit speedily or speeded up. So that thinking about what makes a good mathematician; the thinking about the tasks before you give them to the group; and thinking about the criteria, because I think all those are valuable routes to eventually being able to moderate the task.* Mathematics teacher, Black et al.2011, p.463

> *But I think it would be essential if everybody had clear training and . . . how the portfolio would look, what the tasks . . . would look like. Obviously samples, portfolios you would want, wouldn't you. You would get a sense of what, what task would be appropriate . . otherwise you are going to get teachers going 'I don't know what I'm supposed to do.'* English teacher, Black et al., 2011, p.463

In addition to the two journal articles which describe this work, a short booklet for teachers has also been produced (Black et al., 2013).

This account of the work of one particular project is included here because it helps to illustrate, for the purpose of the main argument of this paper, the general conditions which should be met by any system which is designed to produce valid summative assessments of a student's work. Validity demands can only be met if schools' own assessments are the basis for their summative results, and these in turn can only be valid and meet the requirements of inter-school comparability and trustworthiness by procedures similar to those described above. The work was however incomplete; in particular, it was not able to explore, within its limited time and resources, any possible measures of the reliability of the summative assessments which were produced. Such measures would be essential if any state were to invest in the resources required to develop summative assessments based on teachers' own judgments.

3.4 The status of schools' own assessments in different state systems

This section will give a brief survey of practices in several different country and state systems, across which schools' own summative assessments are given a status which ranges from high to negligible. Whilst this brief survey illustrates some of the lessons learned in the work described above, direct comparisons are of limited scope. All such systems are constructed to meet the need to make the education accountable to the society which both supports and depends on them: they differ widely in the methods they adopt to achieve this aim.

Some of the principles were highlighted in the US National Research Council's study (Pellegrino et al. 2001), which emphasized the need for multiple measures to "enhance the validity and fairness of the inferences drawn" (p. 255), but also stressed that if classroom assessments were to be widely used, careful scrutiny would be needed to ensure that they were adequate in terms of their validity, reliability and fairness. Yet in the USA at present external accountability tests, usually using multiple-choice question, are dominant in most states and as a consequence, teacher judgment for summative assessment in the USA had often been 'found wanting' (Brookhart , 2011). Similar difficulties have been described by Webb (2009), in a report on a USA project to improve the assessment practices of mathematics teachers. In their design and selection of questions and tasks, their interpretations of students' responses and their feedback to students, the teachers were not guided by clear concepts about the role of assessments in classroom learning.

Attempts to broaden the scope of assessments in three states by promoting the use of student portfolios encountered several problems, including weak guidelines for the inclusion of evidence in portfolios, and inadequate training in marking. However these problems have been addressed, in part by inter-school moderation which has been valued also because it has been found to improve school practice (Koretz, 1998; Shapley & Bush, 1999).

For Scotland, implementation of similar ways to develop teachers' summative assessment skills were supported by the absence of state-wide tests, but then local district authorities used their own formal tests as they felt it necessary to have quantitative data for reporting on schools. The study of Boyd and Hayward (2010) found there was an urgent need to improve teachers' assessment literacy and Gardner et

al. (2010) judged that the system was under-designed.

For England, a plan – drawn up in1998 by a group appointed by the Minister of Education – for a new system of national assessment recommended that these be based on a combination, of externally set and marked tests, with assessments of a wider range of each students' own work by their teachers and checked by inter-school moderation. However, that minister was later replaced and the recommendations were rejected (Black, 1997). Gardner et al. (2010) described the narrowing effect of some of the mandatory national tests whilst Fairbrother's (2008) analysis of the mark schemes of such tests showed that the test items called for one-line statements or for box-ticking, each earning either 1 mark or zero, which overall accounted for over 80% of the marks. More recently, the inclusion, in national tests of science, of direct evidence of students 'hands-on' work in practical work has been removed, with teachers required only to report that each candidate has carried out a specified list of such experiments.

A report of a conference organized by the UK Royal Society of Arts in 2017 stated that, for assessment of the A-level (post-16) courses in schools in England concluded that:

> *there was wide consensus that the A level system is outdated and we need to come up with ways to make it broader*

The report went on to recommend:

> *. . . .that teachers have an increased role in assessing student achievement in public qualifications. Many of the world's successful education systems (Finland, New Zealand, Singapore and Ireland) entrust teachers with greater responsibility for assessment, with school-based performance assessments often helping to improve teaching*

One of the report's authors called for a campaign which :

> *. . . . focuses the public and professional debate about education on its highest purposes – like personal fulfilment, societal progress and human flourishing – rather than the proxy goals of tests, targets and league tables (still less the tactics for passing, hitting and climbing them).*
>
> Astle, 2017

Similar points were made in a report in by the UK's Royal Society in 2014: two of their recommendations were:

> *Teachers should have an increased role in assessing student achievement in public qualifications*

and

> *There should be a reduced focus from governments and inspectorates on high stakes accountability measures based on testing*.

A broad range of accounts from different countries was presented in a special issue of the journal 'Assessment in Education' in 2015. Accounts of assessment practices in eight countries were given in separate papers, together with a final overall commentary on these (Black, 2015). Common to all were attempts to promote assessment for learning, but many such attempts had achieved only limited success, with a lack of clarity both about the concept itself and about the relevant procedures entailed. There was wide diversity because of differences in the cultural contexts and because of the different ways in which changes had been introduced. With the sole exception of a report from the USA (Wylie and Lyon, 2015) there was very little information about the classroom discussions which were developed. In one country, teachers had control over all summative assessments of their students, but had not been trained to produce valid assessment tools. The effect of the demands of accountability were a limiting feature in the other countries.

A fuller account of similar issues, including descriptions of the different ways in which they have been addressed in the work of the eight European countries which collaborated in the ASSISTME project, has been presented in a report of that project's finding (Dolin and Evans 2017). However, whilst various components of that account yield useful and relevant insights about the problems involved in enhancing the quality of summative assessments, the project did not attempt to recommend any particular ways to improve the existing systems in the countries involved.

There are indeed some positive examples in several states to offset against the situations described above. In Sweden**,** teachers own assessments determined the summative results for individual pupils. National tests were used for overall calibration of results, with banks of test items available for teachers' use, whilst for school-leaving grades, the teachers results were supplemented with the results of a national aptitude test – in competition for university places the latter have become increasingly important (Wilkstrom and Wilkstrom, 2005).

In several other states there have been state-wide systems to enhance and give status to the summative assessment by teacher and schools. The first example describes the systems which are established in several Australian states, notably New South Wales and Queensland (Stanley et al., 2009): in both these states, the end-of-school 'high-stakes' assessments of students are based on inter-school moderation procedures of each school's assessments using the blind-marking approach outlined above. For the system in New South Wales, formal state tests are taken by all students to serve two purposes: the first is to form 50% of each student's final assessment, the other 50% being based on each school's portfolio assessments. The second purpose is to audit the inter-school moderation process by comparing each school's mean results with the distribution of summative assessment results across all schools. In Queensland, there are no state-wide tests – the results are based solely on the assessments by each student's school. However, moderation meetings between local clusters of schools serve to develop and ensure the inter-school comparability and quality of the assessments. There is a multi-level process whereby, after a local cluster has agreed results, samples from that cluster's schools are then submitted to a further blind-marking procedure in inter-cluster meetings. Wyatt-Smith at al. (2010) give a detailed account of the group moderation process. Extensive investment, extending over about eight years, in training of all teachers in the in-school and inter-school practices was found necessary: one outcome has been that teachers have used the broader scope of the assessment procedures to ensure that these make a direct contribution to students' own learning.

Other work to enhance teachers' summative assessment in New Zealand is described by Hipkins and Robertson (2011) who commented:

> *Whilst teachers have always made judgments informally, moderation as an organised process requires making collaborative decisions to reach consensus agreement, and hence has become an important professional responsibility for all New Zealand's primary teachers.* p.5

Reports of similar positive effects on teaching and learning emerged from the initiative in Canada (DeLuca et al., 2015). One of the teachers involved said 'I have completely changed my style of teaching' and another that :

> *it's not just about sharing success criteria and learning goals, it's now about how we are teaching'.*

There are three general features which emerge from the reforms, interventions and evaluation studies which are discussed above. The first is that it is possible to so develop teachers' summative assessments that they can command public trust. The second is that teachers involved in such work have found that it makes a strong and welcome contribution to their professional development and to their development of students' learning. The third is that these reforms need training and support, sustained over several years,

if they are to succeed .

4 CONCLUSIONS

The main aim of this article is to argue for the recognition of the importance of teachers' assessments and of the need for their development so that teachers can play a central role in the assessments of their students at all levels, from the day to day use of formative responses to guide students' learning to the high-stakes summative assessments which guide students' decisions and progress when they leave school. To achieve this, it is essential that external tests be given either a limited role, or play no direct role, in the state oversight of the achievements of individual students and of schools.

Where teachers' judgments make no contribution to national or state assessment results, the accountability pressures limit their control over their own teaching methods, lower their status, deprive them of full ownership of their work, and undermine the development of their own skills in assessment. If they are trained and supported in attaining full and effective ownership of their work, teachers are far better able to guide all aspects of their students' development, and in particular to equip them to help their students to grow as independent and responsible learners, well able to deal with the new demands they will meet beyond the school.

The central importance of this argument is illustrated by the following statement by the head-teacher of a school in England:

> *(Parents) know that, for the child, the encounter with the teacher is the first major step into outside society, the beginning of a long journey towards adulthood in which the role of the teacher is going to be decisive.* Milroy, 1992, p.57

> *Teachers are, therefore, not in the first instance agents either of the National Curriculum Council (or whatever follows it) or of the State. . . . . The role of the teachers is to attract them progressively into the many realms of the culture to which they belong. This culture consists partly of a heritage, which links them to the past, and partly of a range of skills and opportunities, which links them to the future.*
>
> Milroy, 1992, p.59

References

Abrahams, I, Reiss, M J and Sharpe, R M 2013 The assessment of practical work in school science. *Studies in Science Education* **49**(2), 209–251

Alexander, R. (2008). *Towards dialogic thinking: rethinking classroom talk.* (4th Ed.) York, England: Dialogos .

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

ASSISTME (2017) *Assess Inquiry in Science, Technology and Mathematics Education.* Details available on: assistme.ku.dk/project

Astle, J. (2017) *The Ideal School Exhibition: rediscovering education's true purpose*. Availabe at: https://www.thersa.org/discover/publications-and-articles/rsa-blogs/2017/11/the-ideal-school-exhibition

Bell, B and Cowie, B. (2001) The Characteristics of Formative Assessment in Science Education *Science Education* **85** 536–553.

Black, P.J. (1997) Whatever Happened to TGAT ? pp.24-50 in Cullingford, C. (ed.) *Assessment vs.*

Evaluation Cassell : London

Black, P. (2015) Formative assessment – an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice*, 22(1), 161-177

Black, P (2017) Assessment in Science Education . Ch.22 pp. 295-309 in K.S.Taber and B. Akpan (eds.) *Science Education; An International Course Companion.* Rotterdam: Sense Publishers. ISBN: 978-94-6300-747-4(pbk,) ISBN: 978-94-6300-748-1(hbk.)

Black, P. and Wiliam, D. (1998) Assessment and Classroom Learning. *Assessment in Education: Principles, Policy and Practice*, **5**(1), 7-74.

Black, P. and Wiliam, D. (2009) Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, **21**(1), 5-31.

Black, P., Harrison,C., Lee, C., Marshall, B. and Wiliam, D. (2003) *Assessment for Learning: Putting it into practice.* Maidenhead: Open University Press

Black, P., Harrison, C., Hodgen, J., Marshall, B. and Serret, N. (2010) Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy and Practice*,**17**(2), 215-232.

Black, P., Harrison, C., Hodgen, J., Marshall, M. and Serret, N. (2011) Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice*,**18**(4), 451-469.

Black, P., Harrison, C., Hodgen, J., Marshall, B. and Serret, N. (2013) *Inside the Black Box of Assessment: Assessment of learning by teachers and schools* London: GL Assessment.

Black,P., Wilson, M. and Yao, Shih-Ying. (2011). Road Maps for Learning: A guide to the Navigation of Learning Progressions. *Measurement* **9** (2-3), 71-123.

Blatchford, P., Baines, E., Rubie-Davies, C., Bassett, P., and Chowne, A. (2006). The effect of a new approach to group-work on pupil-pupil and teacher-pupil interaction. *Journal of Educational Psychology*, **98,** 750–765.

Boyd, B., and Hayward, L. (2007). *Exploring assessment for accountability*. Research paper produced for the Assessment is for Learning programme. Accessed Oct. 2009 from http://wayback.archive-it.org/1961/20100730134148/http://www.ltscotland.org.uk/resources/e/genericresource_tcm4579389.asp?strReferringChannel=assess

Brown, M., ed. 1992. *Graded assessment in mathematics (GAIM).* Walton on Thames, UK: Nelson

Bruner, J. (1966) *Toward a Theory of Instruction*. New York: Norton for Harvard University Press

Brookhart, S.M. (2011) *The use of teacher judgment for summative assessment in the United States: Weighed in the balance and (often) found wanting.* Oxford, UK: Oxford Centre for Educational Assessment. http://oucea.education.ox.ac.uk/events/teacher-judgment-seminar/papers-2/

Butler, R. (1988) Enhancing and undermining intrinsic motivation; the effects of task-involving and ego-involving evaluation on interest and performance, *British Journal of Educational Psychology*. **58**(1) 1–14.

Chi, M.T.H. (2009) Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities pp.73-105 in *Topics in Cognitive Science 1*, Cognitive Science Society. Mahwah, N.J: Erlbaum

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd edition) (pp. 443–507). Washington, DC: American Council on Education.

Coffey, J.E.,.Hammer, D., Levin, D.M. and Grant, T. (2011) The Missing Disciplinary Substance of Formative Assessment. *Journal of Research in Science Teaching* **48**(10),1109-1136.

Crooks, T.J., Kame, M.T., and Cohen, A.S. (1996) Threats to the valid use of assessments. *Assessment in Education: Principles, Policy and Practice* **3**(3), 265-285.

DeLuca, C., Klinger, D., Pyper, J., and Woods, J. (2015). Instructional rounds as a professional learning model for systemic implementation of assessment for learning. *Assessment in Education: Principles, Policy & Practice*, **22(1)**, 122–139.

Dolin, J., and Evans, R. (2018) *Transforming Assessment – Through an Interplay Between Practice, Research and Policy* Switzerland: Springer

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality and development.* Philadelphia, PA: Psychology Press.

Fairbrother, R. (2008). The validity of key stage 2 science tests. *School Science Review. **89**(329),* 107-114.

Fitzgerald, A. and Gunstone, R. (2012) Embedding assessment within primary school science lessons: A case study - in R. Gunstone & A. Jones (eds.) *Assessment in Science Education*. Melbourne: Springer

Gardner, J., Harlen, W., Hayward, L. and Stobart, G. (2010) *Developing teacher assessment.* Buckingham: Open University Press.

Greene, J. A., & Azvedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, **77**(3), 354–372.

Hadenfeldt, J. C., Neumann, K., Bernholt, S., and Liu, X. (2016). Students' progression in understanding the matter concept. *Journal of Research in Science Teaching*, **53**(5), 683–708.

Hallam, S. and Ireson, J. (1999) Pedagogy in the Secondary School. Ch.4 pp.68-97 in Mortimore, P. (ed.) (1999) *Understanding pedagogy and its impact on learning.* London: Paul Chapman.

Harlen, W. (2012) On the relationship between assessment for formative and summative purposes in Gardner, J. (ed.) *Assessment and Learning* pp. 87-102 London: Sage

Harrison, C. (2014) Assessment of Inquiry Skills in the SAILS Project *Science Education International* **25**(1), 112-122

Harrison, C., Constatinou, C.P., Correia, C.F., Grangeat, M., Hahkioniemi, M., Livitzis, M., Nieminen, P., Papadouris, N., Rached, E.,Serret, N., Tiberghien, A., and Viiri, J. (2017) Assessment-On-the-Fly: Promoting and Collecting Evidence of Learning Through Dialogue. pp. 83-107 in Dolin, J., and Evans, R. (2018) *Transforming Assessment – Through an Interplay Between Practice, Research and Policy* Switzerland: Springer

Hipkins, R., and Robertson, S. (2011). *Moderation and teacher learning: What can research tell us about their inter-relationships?* Wellington: New Zealand Council for Educational Research.

Johnson, P. and Tymms, P. (2011) The Emergence of a Learning Progression in Middle School Chemistry. *Journal of Research in Science Teaching* **48**(8), 849-877.

Koretz, D. 1998. Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education: Principles, Policy and Practice* **5**(3) 309–34.

Lighthall, F. F. (1988). An organization watcher's view of questioning and discussion. pp. 135–153 in J. T. Dillon (Ed.), *Questioning and discussion: A multidisciplinary study* New York, NY: Ablex.

Mercer, N., Dawes, L., Wegerif, R., and Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal*, **30**(3), 359–377

Milroy, D. (1992). Teaching and learning: What a child expects from a good teacher, pp.57-61 in *Education: Putting the Record Straight.* Stafford, UK: Network Educational Press.

Morell, L. Collier, T., Black, P. and Wilson, M. (2017) A Construct-Modeling Approach to Develop a Learning Progression of how Students Understand the Structure of Matter *Journal of Research in Science Teaching* **54**(8) 1024-1048.

Newton, P. E. (2012)). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, **10** (1-2), 1–29.

Pellegrino, J.W., Chudowsky, N. and Glaser, R. (2001) *Knowing what students know: the science and design of educational assessment* (Washington, DC, National Academy Press).

Pellegrino, J.W., DiBello, L.V. and Goldman, S.R. (2016) A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments, *Educational Psychologist*, **51**(1) 59-81.

Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes. Towards a wider conceptual field. *Assessment in Education: Principles, Policy and Practice*, **5**(1),

85–102.
Poehner, M.E., and Lantolf., J.P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research* **9**(3) 1–33.
Royal Society (2014) *Vision for science and mathematics education : Judging success in education* Available on : https://royalsociety.org/topics-policy/projects/vision/judging-success/
Royal Society of Arts (2017) *The Ideal School Exhibtion: 113-page report on an Inquiry into School Assessments.* Available on: https://www.thersa.org/globalassets/pdfs/reports/rsa-the-ideal-school-exhibition.pdf
Ruiz-Primo, M.A. and Furtak , E.M. (2007) Exploring Teachers' Informal Formative Assessment Practices and Students' Understanding in the Context of Scientific Inquiry *Journal of Research in Science Teaching* **44**(1), 57-84
SAILS (2016) *Strategies For Assessment of Inquiry Learning in Science*. Full details available on www.sails-project.eu/index.html
Shapley, K.S., and Bush, M.J. (1999). Developing a valid and reliable portfolio assessment in the primary grades: Building on practical experience. *Applied Measurement in Education* **12** (2), 111–132.
Shavelson, R. J., Young, D.B., Ayala 1, Carlos C., Brandon, P.R., Furtak, E., Ruiz-Primo, M., Araceli, M.,Tomita, M.K. and Yin, Y. (2008) On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers. *Applied Measurement in Education*, **21**(4) 295-314
Smith, G.A. 1978. *JMB Experience of the Moderation of Internal Assessments*. Manchester: Joint Matriculation Board. Accessible via https://openlibrary.org/books/OL18165228M/JMB_experience_of_the_moderation_of_internal_assessments.
Stanley, G., MacCann, R., Gardner, J., Reynolds, L. and Wild., I. (2009). *Review of teacher assessment: Evidence of what works best and issues for development.* Oxford: Oxford University Centre for Educational Development; report commissioned by the QCA. Accessed Nov.2017 at http://oucea.education.ox.ac.uk/?s=Review+of+teacher+assessment
Webb, D. C. (2009). Designing professional development for assessment.*Educational Designer*. **1**(2). Available on : http://www.educationaldesigner.org/ed/volume1/issue2/article6/
Wiliam, D. 1998. Construct referenced assessment of authentic tasks: Alternatives to norms and criteria. Paper presented at the 24th Annual International Association for EducationalAssessment (IAEA) Conference, May, in Barbados, West Indies.
Wilkstrom, C. and Wilkstrom, M. (2005) Grade inflation and school inflation and school competition: an empirical analysis based on Swedish upper secondary schools. *Economics of Education Review* **24**(3), 309-322.
Wood, D. (1998) *How children think and learn: the social contexts of cognitive development* (2nd edn.) Oxford : Blackwell
Wyatt-Smith, C., Klenowski, V. and Gunn, S. (2010) The centrality of teachers' judgment practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy and Practice* **17**(1), 59-75.
Wylie, E. C., and Lyon, C. J. (2015). The fidelity of formative assessment implementation: Issues of breadth and quality. *Assessment in Education: Principles, Policy & Practice*, **22(1)**, 140–160.